Distribution of traffic accidents

Author: Raúl Fuentes del Pino.
Advisor: Miquel Montero Torralbo
Facultat de Física, Universitat de Barcelona, Diagonal 645, 08028 Barcelona, Spain

Abstract: The aim of this project is to analyze the distribution of traffic accidents in Barcelona, for 10 years of data. The principal motivation is to try to see if there are some patterns in human actions, as has been demonstrated in other studies for other actions like email and letter communications. We will focus our project to approximate this behavior by Poisson process and work with some sampling like the day and night. Also, we will try to observe some human habits. We realized that the approach was quite good, however it is described with two different dynamics, day and night.

I. INTRODUCTION

Social dynamics have been very important over time. There has always an interest in the behavior of groups when they interact individually. We assume that everyone is influenced by others behavior.

Nowadays, social dynamics, is getting more relevant because we have a lot of information about it. There are many factors that determine the actions of every person, so it is impossible to do accurate prediction of the behavior. However, we can observe some patterns and periodicity. The timing of the human actions is highly important, and it is a good magnitude to quantify and has high scientific interest. Human dynamics are described by a function of time and there are many studies that provide an evidence of that. For example, web browsing, email– and letter–based communications, library and stock trading [1]. Individually human actions are well approximated by Poisson process [2].

We are going to study the traffic accidents in Barcelona, see their properties and try to determine the timing of human's actions in accidents and look for other patterns. We will use Guàrdia Urbana open data from 2010 to 2019 [3]. As Guàrdia Urbana tell us, the hour of accidents is annotated when they arrive to the accidents and then people who have suffered the accidents tell them an approximation of it. Ten years of data with more than 10.000 accidents for year, it is enough data to apply some data aggregation like the days of week and we still having enough data.

II. POISSON DISTRIBUTION

We will assume that traffic accidents are well described by a Poisson process. Each human activity is independently form each other, with constant rate λ. Poisson distribution have all the moments with a finite value.

The precision of our data is one hour, so we will count the time lapse between two events and check if it follows this exponential form:

$$\mathrm{P(t)} = \lambda \mathrm{e}^{-\lambda t}. \qquad (1)$$

The time average will be 1/λ, where we supposed that follows an exponential distribution and we apply the maximum likelihood of our data:

$$\mathrm{L}(\lambda|x_1, x_2, \dots, x_n) = \mathrm{L}(\lambda|x_1)\mathrm{L}(\lambda|x_2) \dots \mathrm{L}(\lambda|x_n),$$

$$= \lambda^n [e^{-\lambda(x_1+x_2+\cdots+x_n)}].$$

Solving the likelihood function when λ is maximum:

$$\lambda = \frac{\mathrm{n}}{\sum_{i=1}^{n} x_i}. \qquad (2)$$

Poisson distribution mean that is hard to find some event with very long time. In this case of sending message with social network, maybe normally you answered every day but sometimes this time will be bigger.

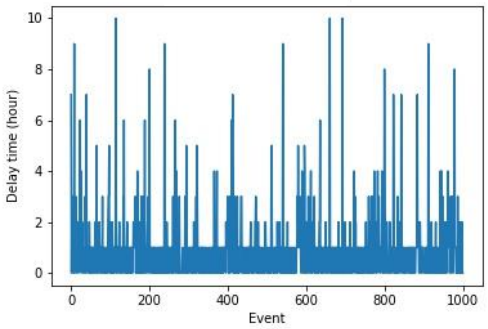

FIG. 1: Example of 1000 events. We can see that are uniformly distributed in time and apparently not periodic. Also, we can observe that there is no large delay time.

If we apply a natural logarithm, we can obtain a linear approach and compare the slope in different situations.

$$\ln\big(\mathrm{P(t)}\big) = \ln\big(\lambda \mathrm{e}^{-\lambda t}\big) = -\lambda \mathrm{t} + \ln\lambda. \qquad (3)$$

As we can see in the formula (3), the slope and the intercept give the same information in the ideal case. Intercept equals to minus one over slope. We will use μ for the inverse of λ that will have units of hour.

* Electronic address: rfuentde13@alumnes.ub.edu

III. RESULTS

After read all of the data sets, clean the data, prepare the variables and eliminate the outliers we start to analyze.

First of all, we need to check how the data of traffic accidents is distributed. We had assumed that it was exponential following the equation (1). Then for 2010 we obtain figure 2.

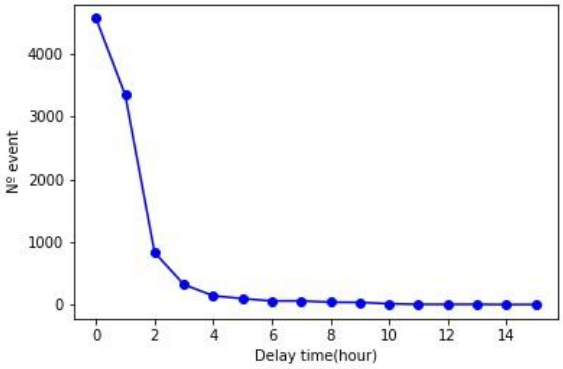

FIG. 2: Number of events in terms of the delay time in hours for the 2010. As we can see, seems to follow an exponential distribution, as the other years.

Another thing that we can observe in figure 2 is that the larger delay time is 15 hours. This value is only one magnitude order bigger than λ, a property of Poisson distribution. The units of λ will be one over hour.

A. Lambda over the years

The next step once we have observed that our data roughly follows an exponential distribution is to check if all the years are the statistical equivalent, then we can group all years as the same sample.

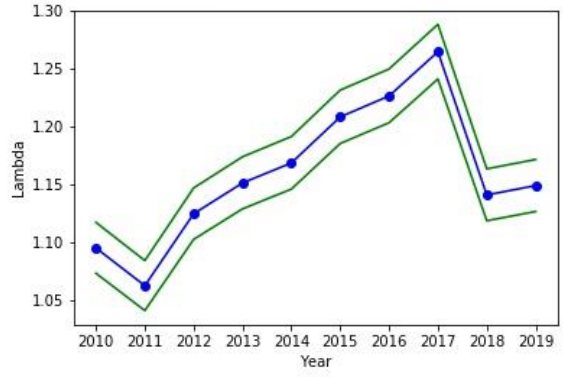

FIG. 3: In blue the λ over the years, in green the variance of λ.

As we can observe, the error in λ for every year cover the neighboring years. Data is not stationary, but their propriety oscillate smoothly.

Next, we compare the λ values, obtained by the three methods described above. If our assumptions were correct, these estimations should agree. As we can see in Table I, this is not the case. This means that we must revise our hypothesis to find the origin of these discrepancies.

	λ_1	λ_2	λ_3
2010	1.0948	0.5302	0.2559
2011	1.0622	0.5367	0.2844
2012	1.1244	0.5397	0.2695
2013	1.1510	0.5559	0.2803
2014	1.1683	0.5680	0.2835
2015	1.2081	0.5593	0.2682
2016	1.2262	0.5486	0.2494
2017	1.2644	0.5506	0.2499
2018	1.1406	0.5529	0.2631
2019	1.1487	0.5599	0.2765

TABLE I: In this table we can observe the different values of λ for every year. The first one is de Maximum Likelihood, the second one is the slope of the equation (3)., and the third λ is obtained with the intercept.

After that, we use the hypothesis testing to compare two exponentials to compare all years and see how much even the distributions are.

FIG. 4: Matrix of correlation of p-value between years.

In figure 4 we observe the p-value for all years. When the value is one mean that is the same distribution. It is the reason because the diagonal is one. We can consider that if the p-value is greater than 0.05, the distribution is the almost the same or similar parameters. Normally all years have the neighboring years with high p-value, it is because the time is continuous and if there not exist a big change that can affect traffic accidents the data distribution will be similar. However, 2017 is different, this behavior was also reflected in figure (3). There was a big change between 2017 and 2018. Also, we can see that years 2018 and 2019 are more similar to those of the period 2012-2014.

B. Biexponential adjust

We started our analysis assuming that the data was distributed in equation (1). However, when we have applied this fit, we saw that our approximation seemed to be inaccurate. This result is observed in table I and in figure 5. When we calculate the regression for each year it looks like figure 5. This characteristic is not an isolated case.

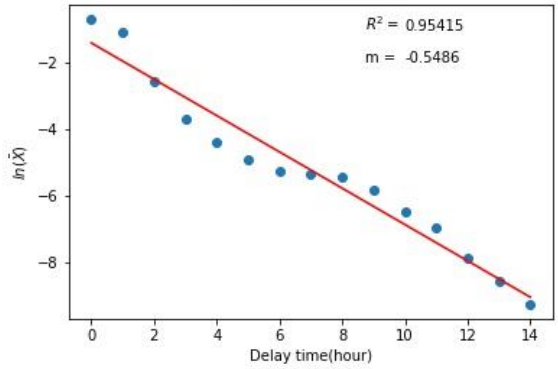

FIG. 5: In y-axis the natural logarithm of the probability (normalized) that one event occurs in function of delay time in 2016. In red the lineal regression of the fit. We added the slope and the R_2.

In this example we can observe what seems a quite good approach to equation (1). However, we can think that it follows a different behavior. In the first part, the lineal regression is above the points, and in the second part, the regression is under the points with an R_2 value of 0.954. One possible solution is to try to adjust a new function with two different dynamics, maybe there are some factor that produce that we had two different dynamics.

Finally, we decide to adjust the following equation:

$$P(t) = ae^{-ct} \cdot \chi(d-t) + a \cdot e^{-cd} e^{-bt} \cdot \chi(t-d). \qquad (5)$$

(The pre-factor of the second term of this equation is intended to make the equation continuous.)

In equation (5), function χ is defined by:

$$\chi(x) = \begin{cases} 1 \; if \; x \geq 0, \\ 0 \; if \; x < 0. \end{cases}$$

Parameter *d* gives us information about when occur the change of dynamic. Finally, the inverse of this parameters *b* and *c* give us an idea of how much frequent the traffic accidents are. The relation between these parameters can helps us to determine the origin of this behavior. Also, the relation with other dynamics of human if exist a characteristically time.

To obtain these new parameters we will apply two different methods.

For the first one method, we will use a python [4] library that find the best fit given a function, with the best parameters. With this method we got a really good fit for the data points as we can see in figure 6.

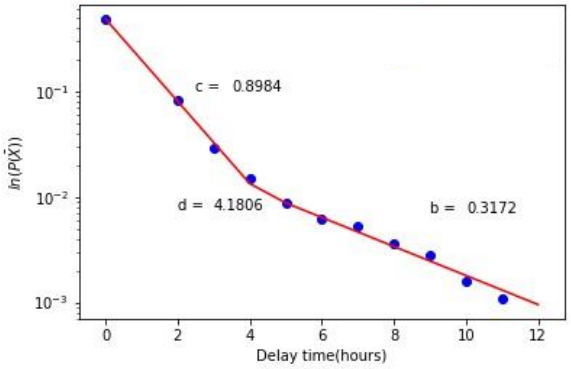

FIG. 6: Probability (normalized) that traffic accident occurs in function of delay time in 2019. The y axis is in log scale.

In this figure we have eliminated the point of one hour of delay time because it can be a introduce and error to the fit. We will give more details about it below.

In this case, we obtain a value of 0.8984 for *c*, it means a μ_C=1.113 and 0.3172 for *b*, with μ_B=3.153. So, after the fourth hour of delay time, the rate change from one to three. In the first interval there is high concentration of accidents.

If we do the same with all years and then the mean of the parameters, we get the following values:

	$\overline{\mu_B}$ (*hour*)	$\overline{\mu_C}$(*hour*)	$\bar{d}$ (hour)
[2010-2019]	1.083±0.004	3.002±0.004	3.90±0.02

TABLE II: Table that shows the mean of the different parameters with the error for the period [2010-2019].

Maybe there is some correlation between these parameters because the value of *d* is similar to sum of the two μ.

The second method is to maximize the product of two coefficient of determination R_2.

The algorithm [5] starts with three points for the first regression and all the others points for the second one. This method consists to create one loop that is adding a new point to the first regression from the second one. For each iteration the algorithm calculates two lineal regressions with the correspondent R_2. We did this process for all the years, to check if every year follow the same dynamics, like happened in the first method, where we want to adjust equation (5) and compare how much different the parameters are.

After this, we calculate the derivate of the product of two R_2 for every year, to determine the maximum value. We did the same process with standard error and we minimize this value. The result obtained with R_2 and with the standard error is the same, both are equal to the fourth hour as we will see in figure 7.

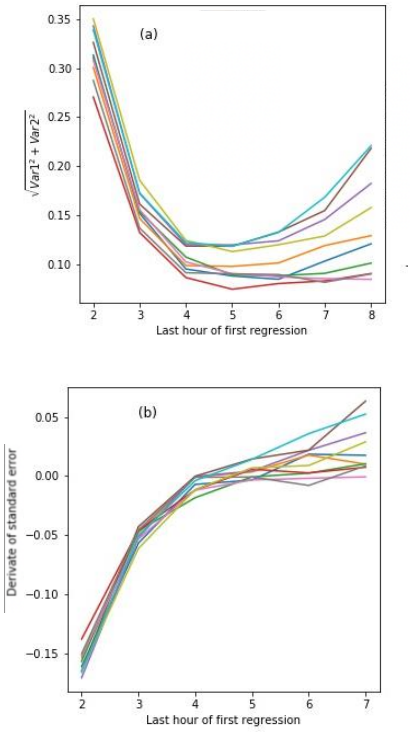

FIG. 7: In plot (a), we can observe how variate de square root of the sum of the two variances at square. In plot (b), the derivate of the that. The X-axis is the last hour of data that the first regression use, is the same for the two plots.

As we can see in this figure, the curve tends to minimize when we have more points, but after some point, it tends to grow up. In the figure 7 (b) we observe that the maximum is when the last hour is the fourth one. The best two fits occur when we create the first lineal regression until the fourth hour.

This result is consistent with the first method, where we obtain that this value is 3.90 (table II).

The value for $\overline{\mu_B}$ using this method is 1.042, is very close to the value obtained with the first one.

With the new approximation (5) that the data follows two different dynamics, the R_2 is 0.986 ± 0.002. This value is better than our first approximation that we obtained an R_2 of 0.95 ± 0.01.

We expected a better coefficient of R_2 because we adjust two lineal regressions.

C. Different data aggregates

Another interesting aspect to analysis is how traffic accidents is affected by the day of the week. The dynamic can be different depending of the day of the week. Maybe, the two dynamics found in the last section can be described because we have two types of day, working day and festive.

When we classify the data and then calculate the same that we did in figure 5, the results obtained are the following

As we expected, we can observe different behaviors. From Monday to Thursday, the results are almost the same, see in table III. However, for Friday we can see a significative increase of accidents. We expected that result because normally on Friday more people use the car.

	λ	μ	R_2
Mon.	-0.7478	1.337	0.960
Tue.	-0.7785	1.285	0.916
Wed.	-0.7651	1.307	0.937
Thu.	-0.8189	1.221	0.952
Fri.	-0.8699	1.150	0.964
Sat.	-0.6545	1.528	0.958
Sun.	-0.6089	1.642	0.976

TABLE III: Table that shows the slope and R_2 value for all days of week, in the period of [2010,2019].

By the other hand, on weekend, the value of rate accidents is higher, people do not need to use the car to go to work. Also, normally people are not in a hurry when they use the car on weekend, so the driving is safer because you can pay more attention.

Another thing to comment is that de R_2 on Sundays is higher than others, possibly because there is no collapse in the circulation, then higher delay time between is more possible. This behavior is more accurate to equation (1).

If we use the same function used in figure 4, which compare two exponential data and give a number of how much equal are, we obtain that Tuesday, Wednesday and Thursday are really similar, also Monday but less than others. For the other hand, Friday, Saturday and Sunday, are totally different. This result agrees what we see in figure 5 for this day, because they have a different slope and maximum range for delay time.

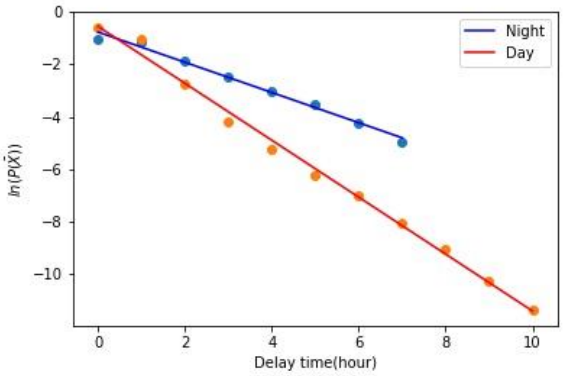

FIG. 8: In Y-axis the ln of the probability (normalized) that one event occurs in function of delay time for all years. In red the lineal regression of the fit during the day hours, the slope is -1.083 with R_2 of 0.994. In blue, the regression during the night [22-05] with a slope of -0.575 and R_2 of 0.989.

After splitting our data with among of the different days of the week we have decided to do the same with day and night. Day hours are defined between 6 and 21 and the night hours are the remainder. As we can see in figure 8, we have obtained a really good results, two different dynamics are observed with a high value for R_2 which is almost one for the day. Also, we can see that during the night, the slope is smaller than in the day, as we expect. The mean of delay time expected that accident occur is two hours during the night and one hour for the day. Also, we can see that the range is higher for the day because the higher possible value for the night is seven.

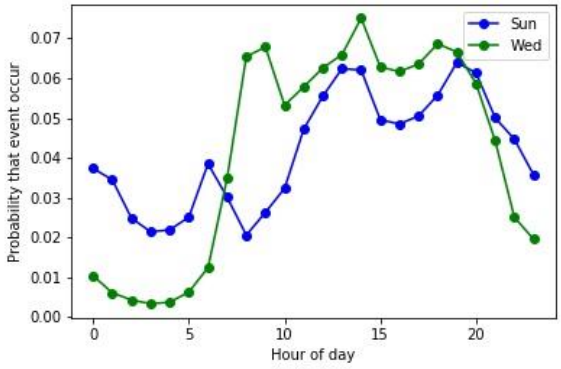

FIG. 9: In Y-axis the probability that one event occurs in function of the hour of the day. The blue line is for Sunday and the green line for Wednesday.

Another interesting thing that data of traffic accidents reflects is that we can see the activity of human actions during the day, like we can see in the figure 9.

On Wednesday, the accidents probability that one accident happens grows up in the period that goes from 6 AM until 10 AM. During this period is when people start to work, and they have to use vehicle. Consecutively, at 10 AM we have a local minimum that may correspond when shops open and the activity of the city become more important. At 2 PM we have the maximum activity, before of going to dinner. Finally, after 6 PM, that is when the activity starts to go down, the probability follows the trend.

Sunday have a peculiarity distribution, different of all days. As we can see, it has a higher probability during the night than other days but also a local maximum at 6 PM, that is the hour when night clubs close. Also, the curve grows up at 8 AM, two hours later than in the other days.

IV. CONCLUSIONS

We analyze the data of traffic accidents over years trying to adjust the best equation. In our first approximation we saw that the distribution is the result of two different behavior. Then we decide to divide our data in working days and festive. With this we get a better approach but not enough good. Finally, when we divide the data in the day and night hours, the result reflected two different regressions with a good value of R_2.

With all of these results, we can say that the principal factor in the traffic accident is the day and night, more than the if the day is festive or not and the day of the week.

When we divide the data in the days of the week, we get a different behavior, but the result was not good enough.

One problem that we had is the uncertainty of the hour when the accident occurs and the poor precision. This was reflected in the delay time of one hour, because the uncertainty is more important for short delay time.

After all, we can affirm that traffic accidents follow an exponential distribution with mean inter-event time about one hour during the daylight hours and about two hours at night.

Acknowledgments

I would like to thank my tutor Miquel Montero Torralbo for his support and guidance during the project. Also, I would like to express my gratitude to Open Data Barcelona for publishing free data sets.

[1] C. Castellano et al, «Statistical physics of social dynamics » *Reviews of modern physics*, vol. 81, pp. 591-646, April-June 2009.

[2] A.Vázquez et al, «Modeling bursts and heavy tails in human dynamics » *Physical review E*, vol. 73, pp. 17, 2006

[3] Open data Barcelona. Guardia Urbana, downloaded at May 2020, https://opendata-ajuntament.barcelona.cat/data/ca/dataset/accidents-gu-bcn

[4] SciPy Libraries, viewed at April 2020, https://www.scipy.org/docs.html

[5] Python code of TFG, https://github.com/Nevvton/TFG